NII_UIT at TRECVID 2021: Video Summarization Task

Khang Dinh Tran[1], Nhat Pham Le Quang[1], Tien Do[1], Tien-Dung Mai[1], An Pham Nguyen Truong[1], Duy-Dinh Le[1], and Shin'ichi Satoh[2]

[1] University of Information Technology, VNU-HCMC, Vietnam
[2] National Institute of Informatics, Japan

Abstract. Finding major life events of a person in video databases is a challenging task. The supervised learning approach requires a annotated dataset to build the life event detector. However, building such dataset is expensive. Our approach to this task is to cast the problem as text matching problem. Face matching is used to locate shots that the queried character appears. We show in the our submitted runs that appropriate fusion of face matching and text matching can answer 2 out of 5 questions on average of the VSUM task.

1 Our Approach

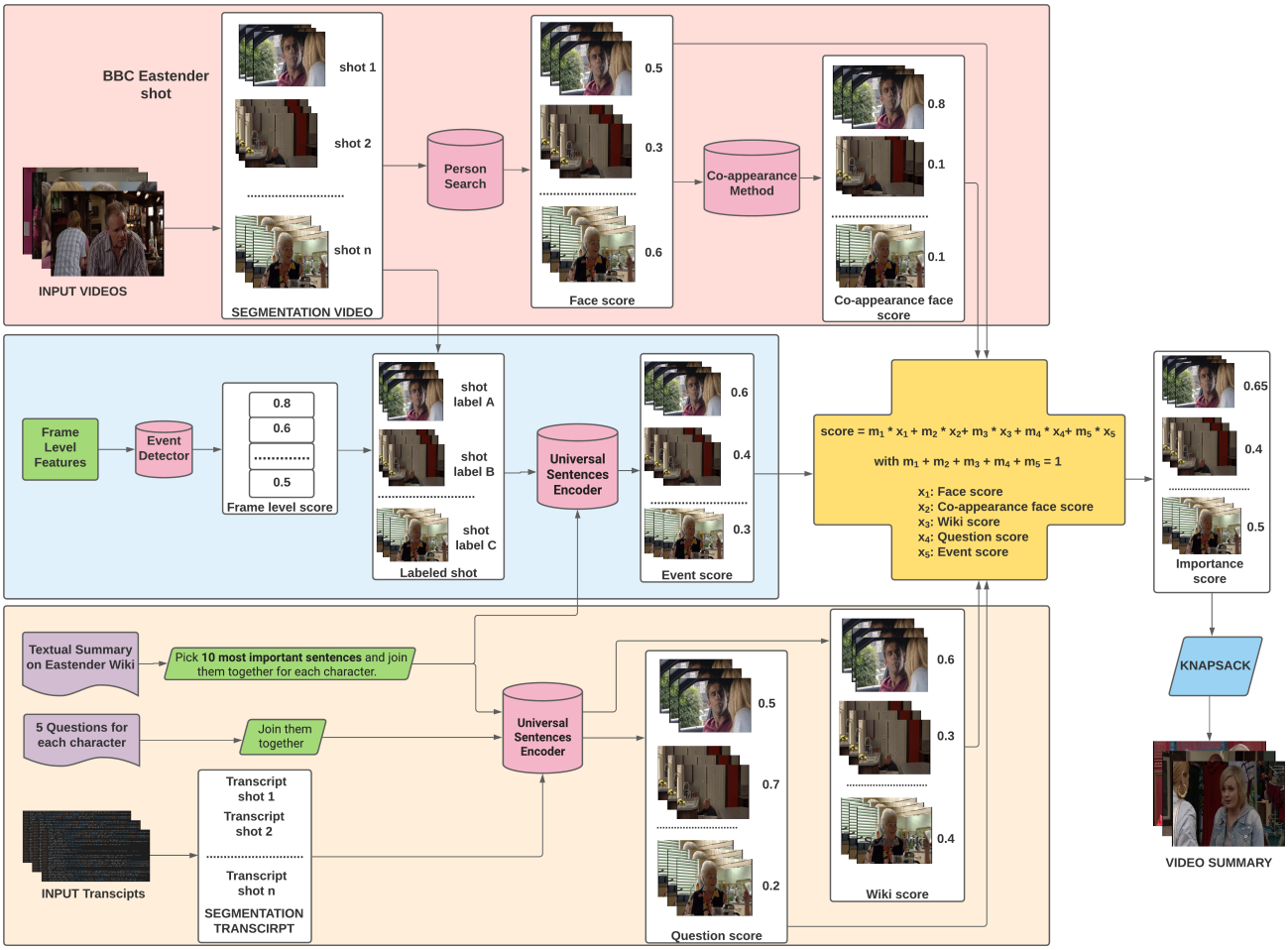

Fig. 1: Our summary framework for TRECVID VSUM task 2021

1.1 Calculating Scores

1.1.1 Face Score We reused the baseline [4] of the NII_UIT team used in the TRECVID VSUM 2020 for finding the queried character. MTCNN [6] is used for face detection, VGGFace2 [2] is used for face representation. An extra step is applied to exclude "bad" looks for faces in query. We used cosine similarity for face matching.

Specifically, we utilized mean-max similarity:

$$sim(query, shot_i) = \frac{1}{N}\sum_{k=1}^{N}(max_{j=1,2,..,M}(cos(desc_k^{query}, desc_j^{shot_i})) \quad (1)$$

where $\cos(A,\ B) = \frac{AB}{||A||.||B||}$ with A and B are the feature vectors of faces in query and faces in shot, respectively. Here, N is the number of examples in the query set and M is the number of faces in the current shot. The variable $desc_k^{query}$ is the descriptor for the $k-th$ face in query, whereas $desc_j^{shot_i}$ is the descriptor for $j-th$ face in $i-th$ shot.

1.1.2 Co-appearance Face Score The major life events of a character often involve with one or several other characters. We used the Co-appearance Face Score to model the co-appearance of the person of interest and another character. The Co-appearance face score of the character A and B is calculated according to the following formula:

$$score_{appearanceAB} = score_{faceA} * score_{faceB} \quad (2)$$

For TRECVID VSUM 2021, we have two groups of characters. The first group has three people who are Jack, Tanya, and Max from video60 to video70, and the second group has two people who are Peggy and Archie from video79 to video89.

For the group includes three characters A, B and C, the appearance score of each character is calculated as follows:

$$score_{appearanceA} = \max(score_{appearanceAB}, score_{appearanceAC}) \quad (3)$$

$$score_{appearanceB} = \max(score_{appearanceAB}, score_{appearanceBC}) \quad (4)$$

$$score_{appearanceC} = \max(score_{appearanceAC}, score_{appearanceBC}) \quad (5)$$

And for the group of two characters A and B, the appearance score of each character is calculated as follows:

$$score_{appearanceA} = score_{appearanceB} = score_{appearanceAB} \quad (6)$$

1.1.3 Text Matching Score From what we have learned from TRECVID VSUM 2020[3], we found that using only visual information is not effective enough to find the correct segments. Therefore, we used a Text Matching method that calculates the similarity of transcript content with semantic paragraphs of text using embeddings from the Universal Sentences Encoder [3]. Specifically, we used cosine similarity to compare the similarity of a transcript of each shot with each other type of semantic text to generate 2 different types of scores, including Wiki and Question.

1.1.3.1 Matching Using Wiki Content

The wiki is a textual summary of a character's life for all the episodes, so we assumed it can describe the main events in that character's life. We collected this data on Eastender Wiki[4] , then picked the ten most important sentences and joined them together for each character.

1.1.3.2 Matching Using Provided Questions (Subtask Only)

For the Subtask, we used the provided questions to search segments closer to the topic. We concatenated the five questions to made the corresponding question content for each character.

1.1.4 Virtual Event Besides finding major life events in segments contained in text data sources such as transcripts, we would like to find them in image data. So we proposed a method using an Event Detection model to detect potential events present in each shot, then saved their labels as text and concatenated them together. Events with scores below 0.8 are ignored. We used the EfficientNet B4 [5] network to train Event Detection model on "USED: A Large Scale Social Event Detection Dataset". [1]

To be able to use this information, we do the same as the Text matching score section. Moreover, we replaced the transcript content with the textual content of the Virtual Event and compared the similarity with the Wiki content.

[3] https://www-nlpir.nist.gov/projects/tv2020/vsum.html
[4] https://eastenders.fandom.com/wiki/Wiki

1.1.5 Importance Score To simplify the selection module, we combined the above score types into an importance score using a linear function as follows:

$$importance\ score = m_1 * x_1 + m_2 * x_2 + m_3 * x_3 + m_4 * x_4 + m_5 * x_5 \tag{7}$$

With constraint:

$$m_1 + m_2 + m_3 + m_4 + m_5 = 1 \tag{8}$$

where:

x_1: Face Score

x_2: Co-appearance Face Score

x_3: Wiki Score

x_4: Question Score

x_5: Virtual Event Score

Since it is impossible to know which type of score is more effective, we choose different sets of $(m_1, m_2, m3, m_3, m_4, m_5)$ parameters at four runs and in both Maintask and Subtask to compare the effectiveness of these types of scores after the results of the organizers. Accordingly, the weighting of the types of scores used is described in the table:

Table 1: The weighting of the types of scores used in TRECVID 2021.

TaskRun	m1	m2	m3	m4	m5
Maintask1	**1.00**	0.00	0.00	0.00	0.00
Maintask2	**0.70**	0.00	**0.30**	0.00	0.00
Maintask3	**0.50**	0.00	**0.50**	0.00	0.00
Maintask4	**0.20**	0.00	**0.70**	0.00	**0.10**
Subtask1	**0.25**	**0.25**	**0.25**	**0.25**	0.00
Subtask2	**0.70**	0.00	0.00	**0.30**	0.00
Subtask3	**0.50**	0.00	**0.25**	**0.25**	0.00
Subtask4	**0.50**	**0.10**	**0.20**	**0.20**	0.00

1.2 Selection

After having importance scores, we selected several shots from all the organizers' segmented shots to create a summary video for each character with four runs with different constraints. To do this, we applied the Knapsack Multiple Constraints problem because the TRECVID 2021 limits both the number of selected shots and the length of the video summary.

Maximize:

$$Z = \sum_{i=1}^{n} u_i s_i \tag{9}$$

With constraints:

$$\sum_{i=1}^{n} u_i l_i \leq L, \quad \sum_{i=1}^{n} u_i \leq K \tag{10}$$

where $u_i = 1$ if shot i is put into knapsack and $u_i = 0$ for the others. K is the number of shots and L is the length of video summary that each run require.

In TRECVID VSUM 2021 Maintask and Subtask, each team have to submit 4 runs for each character following the Figure 2.

2 Experiments

2.1 Dataset

The BBC EastEnders dataset consists of 244 video files, about 464 hours in MPEG4 format. In addition, the data includes transcripts and a small amount of metadata. During this year's mission, we have to create a video summarizing the major life events of each Jack, Max, Tanya character in the range of video60 to video70 and each Archie, Peggy character in the range from video79 to video89. Table 3 provides an overview of the dataset properties.

Character	Max	Jack	Tanya	Peggy	Archie
Start Shot #	shot60_1	shot60_1	shot60_1	shot79_1	shot79_1
End Shot #	shot70_2040	shot70_2040	shot70_2040	shot89_2036	shot89_2036
Images	Images	Images	Images	Images	Images
Max # Shots Run 1	5	5	5	5	5
Max Summary Length Run 1	50 seconds	50 seconds	50 seconds	50 seconds	50 seconds
Max # Shots Run 2	10	10	10	10	10
Max Summary Length Run 2	100 seconds	100 seconds	100 seconds	100 seconds	100 seconds
Max # Shots Run 3	15	15	15	15	15
Max Summary Length Run 3	150 seconds	150 seconds	150 seconds	150 seconds	150 seconds
Max # Shots Run 4	20	20	20	20	20
Max Summary Length Run 4	200 seconds	200 seconds	200 seconds	200 seconds	200 seconds

Fig. 2: Specifies of VSUM Maintask and Subtask 2021[5]

Table 2: Result of our submitted 4 runs on Video Summarization maintask and subtask of TRECVID 2021.

Character	**Run**	**#Shot**	**Summary Length In Maintask**	**Summary Length In Subtask**
Max	1	5	11.7	48.4
	2	10	50.6	58.4
	3	15	74.2	79
	4	20	81.8	149.9
Jack	1	5	11.6	44.2
	2	10	27.8	70.3
	3	15	42.7	96.2
	4	20	70.2	146.1
Tanya	1	5	10.5	43
	2	10	71.5	86.5
	3	15	101.2	83.3
	4	20	106.3	199.6
Archie	1	5	9.4	39.8
	2	10	54.8	76.5
	3	15	75	65
	4	20	30.8	185.2
Peggy	1	5	9	49.2
	2	10	64.8	71.8
	3	15	93.8	87.4
	4	20	24.8	187.7

Table 3: provides an overview of the dataset properties.

Character	Start Shot #	End Shot #	#Shot	Total length (hour)
Jack, Max, and Tanya	shot60_1	shot70_2040	23134	21.04
Archie and Peggy	shot79_1	shot89_2036	20302	19.26

2.2 Result

This year's TRECVID VSUM Task results of the teams are described in Table 4. Our team achieved the highest performance on Run4 Subtask with a surprising figure of 49.2 percentage, much larger than other figures in the table. However, in other Runs, our team did not get the best results, and there was a significant difference in the results of our runs. This disparity is due to our desire to test different sets of parameters when combining score types to evaluate the effectiveness of the score types and figure out how to incorporate them for best performance, as mentioned in the importance score section.

Table 4: Average score of each team in Maintask and Subtask

TeamRun	Percentage in Maintask	Percentage in Subtask
ADAPT1	31.20%	15.60%
ADAPT2	34.20%	11.40%
ADAPT3	27.40%	17.00%
ADAPT4	27.80%	25.00%
EURECOM1	17.40%	32.20%
EURECOM2	30.40%	31.80%
EURECOM3	32.80%	30.80%
EURECOM4	**37.60%**	34.60%
NII_UIT1	7.40%	19.60%
NII_UIT2	12.20%	22.40%
NII_UIT3	29.60%	28.20%
NII_UIT4	22.80%	**49.20%**

An interesting result was found in Table 5. When calculating the total number of distinct questions answered by each team on all runs, all three participating teams answered the same number of 13 out of 25 questions. And there are up to 10 questions all three teams can answer, which raises one question: Are the methods used similar or are these questions more accessible than others.

Table 5: The questions answered by the 3 teams.

CharacterQ	**Question**	**ADAPT**	**EURECOM**	**NII_UIT**
JackQ1	What happens when police break in the door of Jack and Tanya's home?	No	No	No
JackQ2	Where are Max and Jack during the voilent confrontation between them when a gun is drawn?	No	No	No
JackQ3	Who does Jack offer to pay in order to withdraw their statement to the police?	No	No	No
JackQ4	**Why is Jack a suspect in the hit and run on Max?**	**Yes**	**Yes**	**Yes**
JackQ5	What does Jack reveal to Tanya about his dodgy past?	No	Yes	No
MaxQ1	What were the cause of Max's serious injuries which left him in hospital?	No	No	No
MaxQ2	**What is/was the relationship between Max and Tanya?**	**Yes**	**Yes**	**Yes**
MaxQ3	What kind of weapon does Max obtain from Phil?	No	Yes	No
MaxQ4	Where are Max and Jack during the violent confrontation between them when a gun is drawn?	No	No	No
MaxQ5	Who is responsible, or who does Max believe is responsible, for the serious injuries which left him in hospital?	No	No	No
TanyaQ1	What does Tanya reveal to the police while being interviewed at the station?	No	Yes	Yes
TanyaQ2	**What is/was the relationship between Max and Tanya?**	**Yes**	**Yes**	**Yes**
TanyaQ3	**What does Jack reveal to Tanya about his dodgy past?**	**Yes**	**Yes**	**Yes**
TanyaQ4	What does Tanya discover in the sink and on Jack's clothes?	Yes	No	Yes
TanyaQ5	**What big move were Tanya and Jack planning for the future?**	**Yes**	**Yes**	**Yes**
ArchieQ1	**What happens when Phil throws Archie in to a pit?**	**Yes**	**Yes**	**Yes**
ArchieQ2	**What happens after Danielle reveals to Archie that Ronnie is her mother?**	**Yes**	**Yes**	**Yes**
ArchieQ3	Where do Peggy and Archie get married?	Yes	No	No
ArchieQ4	What happens when Archie arrives at the pub after Peggy invited him?	No	No	Yes
ArchieQ5	**What happens when Archie is kidnapped?**	**Yes**	**Yes**	**Yes**
PeggyQ1	Who does Peggy ask to kill Archie?	No	No	No
PeggyQ2	Where do Peggy and Archie get married?	Yes	No	No
PeggyQ3	**Show one of the challenges which Peggy faces in her election run.**	**Yes**	**Yes**	**Yes**
PeggyQ4	What does Peggy overhear Archie saying, which causes their marriage to be over?	No	No	No
PeggyQ5	**What is Janine doing to irritate or anger Peggy?**	**Yes**	**Yes**	**Yes**
		13	**13**	**13**

3 Conclusion

We have proposed a method for finding major life events of a character in BBC EastEnders dataset. We used text matching techniques to search against the transcript. We generated queried sentences using the character's document available from Wikipedia that match with predefined keywords related to life events. The appropriate weights in fusion of face maching score and text matching score can lead a reasonable performance.

References

1. Ahmad, K., Conci, N., Boato, G., VNatale, F.: Used: a large-scale social event detection dataset pp. 505–520 (2016)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
3. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 169–174. Association for Computational Linguistics, Brussels, Belgium (Nov 2018)
4. Le, D.D., Vo, H.Q., Nguyen, D.M., Do, T.V., Pham, T.L.G., Vo, T.L.M., Nguyen, T.N., Nguyen, V.T., Ngo, T.D., Wang, Z., Satoh, S.: NII_UIT AT TRECVID 2020. (2020)
5. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019)
6. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (Oct 2016)